A Novel Schema-Oriented Approach for Chinese New Word Identification

Zhao Lu
Dept. of Computer Science and Technology, East China Normal University
zlu@cs.ecnu.edu.cn

Zhixian Yan
Samsung Research America, Silicon Valley, USA
zhixian.yan@samsung.com

Junzhong Gu
Dept. of Computer Science and Technology, East China Normal University
jzgu@cs.ecnu.edu.cn

Abstract

With the popularity of network applications, new words become more common and bring the poor performance of natural language processing related applications including web search. Identifying new words automatically from texts is still a very challenging problem, especially for Chinese. In this paper, we propose a novel schema-oriented approach for Chinese new word identification (named "ChNWI"). This approach has three main steps: (1) we suggest three composition schemas that cover nearly all two-character up to four-character Chinese word surfaces; (2) we employ support vector machine (SVM) to classify Chinese new words of three schemas using their unique linguistic characteristics; and (3) we design various rules to filter identified Chinese new words of three schemas. Our extensive evaluations with two corpora (Chinese news titles and CIPS-SIGHAN 2012 CSMB) show ChNWI's efficiency on Chinese new word identification.

1 Introduction

With the rapid development of information technology, as well as the growth of social networks (e.g., Chinese Microblog, WeChat), Chinese new words are constantly being created and their usages have become an inevitable phenomenon. Automatic identification of new words plays an important role in a number of areas in Chinese language processing, such as automatic segmentation, information retrieval and machine translation (Zhang et al., 2010; Duan et al., 2012). In the Chinese new word identification (NWI) task, new words refer to new composition words that are not registered in the dictionary of a Chinese segmenter.

Statistical approaches are the most widely used methods in NWI. The previous methods extract some linguistic features of new word compositions, i.e., *word composition probability*, *co-occurrence probability*, *mutual information*, and *word frequency*, while they assume above linguistic features playing the same impact on various word surfaces (Chang and Lee, 2003; Li et al., 2008; Zhang et al., 2010). Some methods also have binary decision, either "new words" or "not new words". A SVM-based method (Li et al., 2008) aims at two word surfaces, NW11 and NW21, and the method uses same linguistic features for the two surfaces. Other statistical models, for instance, a latent discriminative model (Pang et al., 2009), a linear-time incremental model (Zhang et al., 2012) and conditional random fields (CRFs) model (Wang et al., 2012), are designed for NWI.

Recently, some hybrid methods have been suggested. These hybrid methods employ more or fewer rules for statistical methods to obtain an optimal efficiency of identification. However, the rules these methods used are created by the people, which cause these methods are not suitable for other new word composition schemas (Zhang et al., 2006; Jiang et al., 2011; Xi et al., 2012).

Despite the wide studies of new word identifications, accurately identifying Chinese new words from texts automatically is still a very challenging task because of the following reasons:

- Most existing studies focus on English and these methods are not suitable for Chinese. Chinese new words have less morphology variations than many other languages, and there is a lack of capital clues as in English. In Chinese, there are not special symbols implying boundaries between two words and any adjacent characters can form a word. This is one main reason of the difficulty to

27th Pacific Asia Conference on Language, Information, and Computation pages 108－117

recognize new words from texts.

- A survey of the literature indicates that there are eleven surfaces of four-tuple Chinese words, while those methods focus on two surfaces (i.e., NW11 and NW21). They use same linguistic characteristics and same filtering rules for the two surfaces. The two aspects cause the lower accurate rate and the problem of data sparseness.

To address these challenges, in this paper, we propose a schema-oriented Chinese new word identification approach which combining SVM and rules, it is called "ChNWI". The ChNWI approach has two main parts, i.e., (1) ChNWI training process, in which we first define three word composition schemas, their particular linguistic characteristics, and one basic feature model with other three feature models for three schemas; (2) ChNWI testing process, in which we identifying new words of three schemas from segmented fragments using various filtering rules of three schemas. Concluded, this paper has the following three main contributions:

- We classify eight of eleven surfaces of four-tuple Chinese words into three composition schemas, i.e., single-character schema, affix schema and NW22 schema. We study their special linguistic characteristics of the three schemas.

- We design a rich set of features models for the three schemas by analyzing their linguistic characterises. We hereinafter apply SVM as our basic classifier due to its robustness, efficiency and higher performance than other classifiers, for instance, Perceptron, Naive Bayes and kNN (Li et al., 2008). Furthermore, we design filter rules for the three schemas to refine the NWI decision.

- We evaluate ChNWI on two corpora, *i.e.*, a collected Chinese news title dataset and a popular MicroBlog dataset. The experimental results show the efficiency of ChNWI on Chinese new word identification.

The remaining sections of this paper are organized as follows. Section 2 presents the main framework of our ChNWI approach. Section 3 introduces three new word composition schemas, their linguistic characteristics and their feature models. Section 4 discusses the training process and the test process of ChNWI. We conduct several experiments and analyze experimental evaluations in section 5. Finally, we conclude this paper and discuss future work.

2 Our ChNWI appraoch

In this section, we first formulate the task in this paper, then we present our approach in general.

The task of identifying Chinese new words in this paper is concluded as: after extracting strings of three kinds of schemas from segmented fragments, we compute the confidence degree of these strings using both their special linguistic characteristics, together with SVM; we select these strings with their confidence degree larger than a certain threshold as new word candidates.

The confidence degree of a Chinese new word with the feature set x belongs to the category y is defined as the co-occurrence probability $p(x, y)$ of the category y and the feature set x. The category y refers to "Chinese new words" or "not Chinese new words", and x is the feature vectors of a new word. Formally, given a training sample set, $T = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, where $x_i \in R^n, y_i \in \{-1, 1\}$. x_i refers to the feature vectors of new words, y_i is the category of a new word.

The framework of ChNWI is shown in Figure 1. Two main parts of the suggested ChNWI approach are the training process and the testing process.

The training process includes: (1) we first segment and POS tagging the training corpus using a Chinese word segmenter; (2) After extracting linguistic characteristics of three schemas, we generate three feature vectors for three schemas using their positive samples and negative samples; (3) Three feature models for the three schemas are generated using the SVM classifier.

The steps of the ChNWI testing process are: (1) we segment and POS tagging the test corpus and extract potential strings of three schemas using two suggested algorithms; (2) we extract three feature vectors for three schemas using the extracted linguistics characteristics during the ChNWI training process; (3) we identify new word candidates of three schemas using the three generated SVM models. Finally, we suggest various rules to filter all candidates.

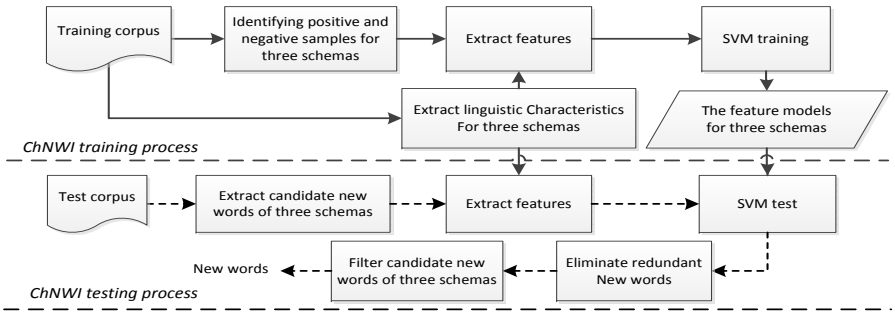


Figure 1: The ChNWI Framework

3 Three schemas of Chinese new words and their feature models

In this section, we present three schemas of Chinese new words and define their feature models.

3.1 Various surfaces of Chinese new words

Literature (Jiang et al., 2011) shows that, all four-tuple Chinese new words can be classified into 11 compositions, i.e., 53 % new words of NW11, 31% new words of NW21, 5% new words of NW12 and NW31, and 11% other schemas. Here, NW is the abbreviation of New Word, 1 refers to a single character, 2 refers to a binary word, 3 refers to a ternary word. After investigating the features of these compositions, we classify four-tuple Chinese new words into three schemas, i.e., *single-character schema*, *affix schema* and *NW22 schema*. The three schemas cover nearly above 11 compositions.

3.1.1 Single-character schema

The new words of *single-character schema* are composed of up to four consecutive single characters. The single-character schema includes, NW11, NW111 and NW1111. Some examples of single-character schema are:

(1) NW11, 蚁/ng 族/n (yi/ng zu/n)

(2) NW111, 经/n 适/n 房/n (jing/n shi/n fang/n)

(3) NW1111, 反/n 独/d 促/v 统/vi (fan/n du/d cuc/v tong/vi).

There are less linguistic characteristics for new words of single-character schema mainly because of most of all single characters have no combined features with their neighboring ones, thus up to four adjacent characters can be viewed as a new word.

3.1.2 Affix schema

The second surface type is “affix schema”. A new word of affix schema is composed by a single character and an existing word. Affix schema can be further classified as *prefix schema* and *suffix schema*. Prefix schema includes NW12 (a single character with an existing binary word) and NW13 (a single character with an existing ternary word), e.g., 反通胀(Anti inflation). Suffix schema includes NW21 (an existing binary word with a single character) and NW31 (an existing ternary word with a single character), for example, 国土部(Ministry of Land and Resources).

Both prefix schema and suffix schema have strong linguistic characteristics. The first character is easy to combine with a binary word to compose a ternary new word, or with a ternary word to constitute a four tuple new word. These kinds of first characters are viewed as *prefix letters*, such as, 零(zero), 软(soft) and 反(anti).

The last character (or the tail character) of suffix schema is easy to combine a binary word to form a ternary new word, or with a ternary word to form a four tuple new word. We view the kinds of tail characters as *suffix letters*, for instance, 部(department), 率(rate) and 式(style).

3.1.3 NW22 schema

New words of NW22 schema are mainly composed by two binary words. Some examples are 人口普查(Census) and 热带风暴(Tropical storm). Unlike single-character schema and affix schema, this kind of new words have less special linguistic characteristics for the reason of two adjacent binary words can compose a new word of NW22. Since there are not significant characteristics of NW22, it is difficult to identify these new words.

3.2 Feature models for three schemas

We first suggest a base feature model for three schemas, then we propose a special feature model for each schema.

3.2.1 Basic feature model

For new words of three schemas, some linguist characteristics are important, i.e., co-occurrence, mutual information, word frequency and adjunct categories. The base feature model (Base$_F$) for three schemas is defined as follows,

$$Base_F\{F_F, F_{COP}, F_{MI}, F_{AV}\} \quad (1)$$

here, F_F refers to word frequency, F_{COP} is co-occurrence probability or average co-occurrence probability, F_{MI} is mutual information or average mutual information, F_{AV} is adjacent categories.

Word frequency is a basic characteristic of new words, especially for NW22 schema. This characteristic is an important aspect of determining whether a string is a new word or not. We view a string S in a corpus is a new word candidate if its frequency is larger than a pre-defined threshold. In this paper, we set the threshold to 2 for the aim of covering mostly new word candidates.

Co-occurrence probability show the tightness degree of two Chinese characters or two words A and B. The higher their co-occurrence probability, the higher the tightness degree of A and B is. The greater the tightness degree is, the more easier A with B to compose a new word.

Mutual information indicates the relevant degree of two continuous strings A and B. Mutual information not only reflects the possibility of the combination of two continuous strings to be a word, but also measures the internal relevant degree of a word. We use *average mutual information* to indicate the coupling degree of continuous characters or words in a string S (Luo and Sun, 2003). The higher the average mutual information of S is, the higher its coupling degree is. Which means the higher possibility of S is to be a new word (Zhou, 2005).

Adjacent category represents the relevant degree among a word (or a string) with its context. Adjacent category $AV(S)$ can be further divided into *left adjacent category* (L_{AV}) and *right adjacent category* (R_{AV}). Given a Chinese string S, its adjacent category is defined as follows,

$$AV(S) = min\{L_{AV}(S), R_{AV}(S)\} \quad (2)$$

here, $L_{AV}(S)$ and $R_{AV}(S)$ refer to the numbers of the words in which the string S appearing in the left or in the right of the words respectively.

In a sentence, a string is viewed as a word if it satisfies that, its cohesive degree is higher and its coupling degree with its context is lower. For a term, its various contexts cause its left adjacent category and its right adjacent category are large numbers. From this consideration, for a Chinese string S, if its left adjacent category value or its right adjacent categories are larger than a predetermined threshold, which means that the string S is loose with its context and it is higher possibility of being a Chinese new word. That is the reason we view the two adjacent categories with lower values to be the adjacent category of the string S in Equation (2).

3.2.2 Feature model for single-character schema

As a new word of single-character schema is a string of continuous characters in a segmented fragment, for single-character schema, we add *independent word probability* (IWP) into the base feature model to get a new feature model, which is called as the feature model of single-character schema (F_{single}) as follows,

$$F_{single}\{F_{IWP}, F_F, F_{COP}, F_{MI}, F_{AV}\} \quad (3)$$

Independent word probability of a string S ($S = c_1, c_2, ..., c_n$) is defined as the joint probability of all characters in the string. We assume that, the higher the independent word probability of a string S is, the higher the probability of S being a new word is. Based on the assumption, we take a string as a new word candidate if its $IWP(S)$ is larger than a pre-defined threshold.

3.2.3 Feature models for affix schema

New words of affix schema have relatively significant linguistic features. That is the probability of the affix characters appearing in the head or the tail of a word is very high. That is to say, the affix characters are easy to compose new words together with existing words or other characters. From this observation, we can compute the

head-character word probability $IWP(C, f)$ and the tail-character word probability $IWP(C, l)$ for a word of affix schema. We further classify the feature model of affix schema into two categories, the prefix feature model (F_{prefix}) and the suffix feature model (F_{suffix}), as follows,

$$F_{prefix}\{F_{IWP}(f), F_F, F_{COP}, F_{MI}, F_{AV}\} \quad (4)$$

$$F_{suffix}\{F_{IWP}(l), F_F, F_{COP}, F_{MI}, F_{AV}\} \quad (5)$$

here, $F_{IWP}(f)$ and $F_{IWP}(l)$ refer to the head-character word probability and the tail-character word probability respectively.

3.2.4 Feature model for NW22 schema

The third schema type is NW22 schema. The new word of NW22 schema is a combination of two existing words. It is obvious that, word probability, head-character word probability or tail-character word probability do not reflect the unique characteristics of NW22 schema. To NW22 schema, both the degree of combination between two existing words and the context of two words are important features. Therefore, we use the base feature model for NW22 schema only.

4 ChNWI training and testing process

4.1 The training process of ChNWI

We first determine positive samples and negative samples for three schemas respectively.

For single-character schema, positive samples mainly refer to words up to four characters in the dictionary of a segmenter, i.e., ICTCLAS, and any substrings of these words are not words. For example, 逃逸(escape), 麦当劳(McDonald's) and 说三道四(make irresponsible remarks) are words in the dictionary, while any sub-strings of the three words are not registered in the dictionary. Negative samples are the extracted continuous strings of NW11, NW111 and NW1111 in segmented fragments, while these strings are not registered as words in the segmenter.

The positive samples of affix schema are ternary words or quaternary words in the dictionary of a segmenter, and parts of these words are words also. For example, the first two characters of the word 户口本(hukou ben) is a word, and the last three characters of the word 喝西北风(he xibeifeng) is a word also. The negative samples of affix schema are these strings combined with a character and a word of NW12, NW13, NW21 and NW31, while they are not words in the dictionary.

For NW22 schema, positive samples are quaternary words in the dictionary, and half of these words are words also. Such as 历史纪录(historical record) and 汉语拼音(Chinese pinyin), parts of the two words are binary words. Negative samples are these strings combined by two binary words while they are not words in the dictionary.

Then we use LibSVM (Chang and Lin, 2011) to gain three SVM models for three schemas using positive samples and negative samples. In order to improve the accuracy of the SVM training model, we manually choose some negative samples as positive sample for three schemas respectively.

4.2 ChNWI testing process

4.2.1 Extracting new word candidates of three schemas

We suggest three methods to extract new word candidates of three schemas respectively.

(1) Extract new word candidates of single-character schema

As we discussed above, a new word of single-character schema is made up of two or more continuous characters in segmented fragments. That is, given a segment fragment $T = \{X_1X_2...X_i...X_n\}(1 \leq i \leq n)$, here, X_i is a word or a character. If there is a string $NW(i, j) = \{X_iX_{i+1}...X_j\}$ in T and each X_i in NW is a character, then we view the string NW as a new word candidate of single-character schema. If the length of NW is larger than 2, then its all sub-strings with lengths larger than or equal to 2 are new word candidates also. If $i = 0$ or X_{i-1} is not a character, and if $j = n$ or X_{j+1} is not a character, then $NW(i, j)$ is viewed as the *longest new word candidate*. For example, both 经适(jing shi) and 适房(shi fang) in 经适房(jing shi fang) are all viewed as new word candidates of single-character schema, 经适房(jing shi fang) is a longest new word candidate.

Given three strings of single-character schema, A, B and C, if there is $A = B+C$, and the lengths of B and C are smaller than or equal to 2, A is viewed as the *parent string* of B and C and both B and C are viewed as two *sub-strings* of A.

We present the process of extracting new word candidates of single-character schema as follows: firstly, we extract the longest new word candidates from the segmented test corpus, and count

their frequencies using Algorithm 1; then, for each longest new word candidates, it's all substrings are extracted and their frequencies are counted using Algorithm 2.

Algorithm 1: The Candidate New Word Detection Algorithm(CND)

```
Input: SSTC
Output: slpuw, spuw
begin
    for each a_i in SSTC do
        get a_i = w[0]w[1]...w[k];
        for each w[j]w[j+1] in a_i do
            if the length of w[j] == 1 then
                if the length of w[j+1] == 1 then
                    add to spuw;
                else
                    N(w[j]w[j+1]) ++;
    set temp to null;
    for each w[j] in a_i do
        if length(w[j]) == 1 then
            w[j] appended to temp;
        else
            If length(temp) > 1 if temp not in slpuw then
                add temp to slpuw;
                set temp to null;
            else
                N(temp) ++;
                set temp to null
```

We use Algorithm 1 to extract all longest new word candidates of single-character schema in a segmented text $SSTC$. Here, $a = \{w[0]w[1]...w[k]\}$ is a segmented fragment in n $SSTC$. w is a part of a, it is maybe a word, a Chinese character, a number or an English character. $N(w)$ is the frequencies of w in the segmented fragments, $length(w)$ is the length of w, $slpuw$ is the longest new word candidate set of single-character schema, $spuw$ is the new word candidate set of affix schema.

Algorithm 2 is a sliding window algorithm which is used to extract all sub-strings of each longest new word candidate and their frequencies. The input and output of Algorithm 2are $spluw$ (the longest new word candidate set) and the new word candidate set. The main idea of Algorithm 2 is to traverse each longest new word candidate using a sliding window algorithm to extract all substrings with their lengths are larger than or equal to 2, and to count their frequencies.

(2) Extract new word candidates of affix schema

Algorithm 2: The Candidate New word Detection Algorithm(CND)

```
Input: slpuw
Output: A set subset of substring
begin
    for each c_k in slpuw do
        let s = c_k, j = 2, substring is null;
        for (; j < length(s); j++) do
            for (i = 0; i+j-1 < length(s); i++) do
                substring = s.sub(i, i+j);
                if substring not in subset then
                    added to subset;
                    N(substring) = N(s);
                else
                    N(substring) ++;
```

All new word candidates of two kinds of affix schema are collected using Algorithm 1 also. The main steps of extracting new word candidates of affix schema are: firstly, we traverse each segmented fragment, collect all strings of NW21 or NW31 as new word candidates and add them into the new word candidate set of suffix schema; then we collect all strings of NW12 or NW13 as new word candidates, and add them into the new word candidate set of prefix schema.

(3) Extract new word candidates of NW22

For new words of NW22 schema, the extraction process is: collect all strings of NW22 schema as new word candidates, and add them into the new word candidate set of NW22 schema also.

4.2.2 Eliminating redundant new word candidates

After extracting all new word candidates of three schemas, we will further eliminate all new word candidates with their frequencies less than 2, and eliminate all redundant new word candidates. For all new word candidates of single-character schema, since we collect all longest new word candidates and their sub-strings as new word candidates, there are redundant candidates in the collection.

The main steps of eliminating these redundant strings are as follows: given a parent string $C_iC_{i+1}...C_{i+j+1}$, its two substrings $C_{i+1}C_{i+2}...C_{i+j+1}$ and $C_iC_{i+1}...C_{i+j}$, the differences between the frequency of $N(C_iC_{i+1}...C_{i+j+1})$ and the frequencies of its substrings, $N(C_{i+1}C_{i+2}...C_{i+j+1})$ and

$N(C_iC_{i+1}...C_{i+j})$, is marked as a. If a is smaller than a predefined threshold b, then we view the two sub-strings are redundant. We remove the two strings and only keep the parent string. On the contrary, if the frequency $N(C_iC_{i+1}...C_{i+j})$ is larger than the frequency $N(C_iC_{i+1}...C_{i+j+1})$, or the frequency $N(C_{i+1}C_{i+2}...C_{i+j+1})$ is larger than the frequency $N(C_iC_{i+1}...C_{i+j+1})$, and the difference between them is larger than b, then we remove the parent string，and keep two sub-strings. In this paper, we set $b = 2$ for the minimum length of Chinese word is 2.

4.2.3 Filtering new word candidates

We design various filtering rules for the single-character schema and the affix schema.

For single-character schema, we use stop words to filter new word candidates. For example, 在(zai), 将(jiang), 称(chen) are often used in texts, while they with other characters or words cannot compose new words. So, for all candidates of single-character schema, if a word starts or ends with these characters, we will eliminate these candidates.

For affix schema, we use a head-character list and a tail-character list for the aim of filtering new word candidates. Some prefix characters, examples including "副, 近, 新" (fu, jin, xin) are often used in prefix schema. During the training process, we have added the top N of characters with their $IWP(f)$ values are bigger into the head-character list. For suffix schema, some suffix characters, for example, " 门, 热, 控" (men, re, kong) are used often. During the training process, we also add the top N characters with their $IWP(l)$ values are bigger into the tail-character list. We design the filtering rules of affix schema as: if the first character in a new word candidate of prefix schema is found in the tail-character list, then we ignore the new word candidate. For example, the first character 案(an) of a new word candidate 案抓获(an zhua huo) is in the tail-character list, so we remove the candidates. Similarly, if the tail character in a new word candidate of suffix schema is found in the head-character list, then we ignore the candidates.

For NW22 schema, there are not special rules to filter new word candidates of NW22 schema.

5 Experimental results and analysis

As we discussed above, new words studied in this paper related to the dictionary of the ICTCLAS segmenter [1], a popular segmenter developed by the Chinese Academy of Sciences. To test the efficiency of our approach, we design three experiments on three corpora. The first corpus is set of the domestic news titles on Sina.com.cn from June 2010 to July 2012, which contains 0.12 Million news titles. We divide the corpus into two parts, one is the testing corpus and the other is the training corpus. The second one is the MicroBlog corpora of CIPS-SIGHAN CLP 2012 Chinese Segmentation on MicroBlog Bakeoff (CSMB) (Duan et al., 2012), which contains 5,000 sentences.

In each ChNWI training process, we use cross-validation method to obtain the optimal training parameters. We divide the training corpus averagely into 10 parts, one is used to verify, the others are used for training. The numbers of features are: four features for single-character schema, six features for affix schema and four features for NW22 schema. The training time of every experiment for ChNWI models is not more than 4 minutes and the testing time is not more than 5 second using a laptop with an Intel(R) Core(TM) i3 CPU and 2.92G RAM.

For evaluation, we adopt the same evaluation method defined in the CSMB bake-off task, precision (P), recall(R) and F-measure.

$$P = \frac{\text{Number of new words correctely identified}}{\text{Number of new words are identified}} \quad (6)$$

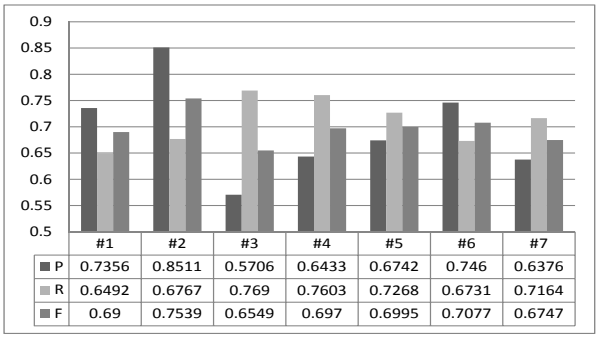
$$R = \frac{\text{Number of new words correctely identified}}{\text{Number of new words in the corpus}} \quad (7)$$

$$F = \frac{2 * P * R}{P + R} \quad (8)$$

5.1 Experiments on the first corpus

In the first experiment, we investigate the related contributions of each feature model of each schema of ChNWI. The experimental results are shown in Figure 2. In Figure 2, #1 and #2 refer to $Base_F$ and F_{single}+Filtering rules of Single-character schema, #3 and #4 are $Base_F$ and F_{prefix}+Filtering rules of Prefix schema, #5 and #6 refer to $Base_F$ and F_{suffix}+Filtering rules of Suffix schema, and #7 refers to $Base_F$ of NW22 schema respectively.

[1] ICTCLAS, http://www.ictclas.org/



	#1	#2	#3	#4	#5	#6	#7
P	0.7356	0.8511	0.5706	0.6433	0.6742	0.746	0.6376
R	0.6492	0.6767	0.769	0.7603	0.7268	0.6731	0.7164
F	0.69	0.7539	0.6549	0.697	0.6995	0.7077	0.6747

Figure 2: Experiments on contributions of various feature models

Test on single-character schema

In the feature model F_{single}, independent word probability is a special linguistic characteristic. To show the effectiveness of independent word probability, we first use the base model, $Base_F$, then we use the feature model F_{single} of single-character schema, together with the corresponding filtering rules.

In Figure 2, #1 and #2 are the experimental results of single-character schema. To some extent, our approach can identify new words of single-character schema effectively. Especially we add the feature F_{IWP} and the relevant filtering rules to the base model, the precision rates improves 11.6% and F-value also increase 6.39%.

Test on Affix schema

Affix schema can be divided into prefix schema and suffix schema.

The feature model for prefix schema is F_{prefix}. In which, the first word probability is an important linguistic characteristics. To show the contribution of first word probability, we first use the base model, $Base_F$, then, we use F_{prefix} with the corresponding filtering rules. The experimental results of prefix schema are shown as #3 and #4 of Figure 2. Our approach has good effectiveness of identifying new words of prefix schema also. Using F_{IWP} and filtering rules, the correct rate improves 7.27%, while F value improves 4.21%.

In the feature model F_{suffix} of suffix schema, the tail word probability is also an important feature. We first use the base model $Base_F$, then we employ F_{suffix} and the relevant filtering rule. The experimental results of suffix schema are #5 and %6 of Figure 2. Similar to prefix schema, our method has better efficiency on identifying new words of suffix schema. After using $F_{IWP(l)}$ and filtering rules, the correct rate improves 7.18% and F-value improves 0.8%.

Test on NW22 schema

As we discussed above, there are less linguistic characteristics of NW22 schema, so we use the base model $Base_F$ as the feature model of NW22 schema. The experimental result of NW22 schema is #7 of Figure 2. #7 shows that our method has better effectiveness on identifying new words of NW22 schema. The F-score of NW22 schema is more than 67%.

5.2 Experiment on MicroBlog Corpora

We perform the second experiment to find how ChNWI improves the performance of a Chinese segmenter. We test ChNWI on the MicroBlog Corpora suggested by CIPS-SIGHAN-2012 CSMB. The corpora includes 294 new words (14%) and 252 rule-based combination of words (12%). Both the two words are unregistered words to a segmenter. The performance of the two test points is, the max correct numbers of the two test points are 65 (22.1%) of new words and near 70 (27.8%) of rule-based combination of words (Duan et al., 2012). Which shows that the systems submitted may not deal with unregistered words well.

The CIPS-SIGHAN-2012 CSMB provides no training set, we train ChNWI on the training corpus used in the first experiment. In the second experiment, we select the ICTCLAS segmenter and the suggested ChNWI is used as post processing. The experimental results are shown in Figure 3 and Figure 4. All data of the maximal (Max) and the average (Avg) performance of Figure 3 and Figure 4 are from the report (Duan et al., 2012).

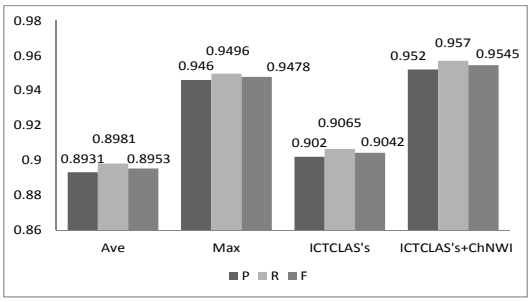


Figure 3: Experimental results of ICTCLAS's with ChNWI on MicroBlog corpus

Figure 3 shows that, compared with ICTCLAS's, F-score of ICTCLAS's + ChNWI is improved near 5%. Compared with Avg and Max,

F-scores of ICTCLAS's + ChNWI are improved 0.6% and near 6 % respectively.

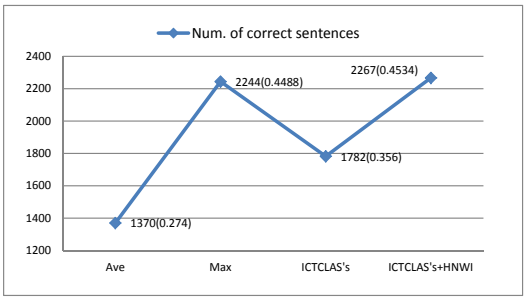


Figure 4: Numbers (and percentages) of correct sentences segmented by ICTCLAS's with ChNWI in MicroBlog corpus

Figure 4 shows the numbers (and percentages) of correct sentences segmented by ICTCLAS's and ICTCLAS's+ChNWI. The number and percentage of correct sentences are improved 485 and 9.7% respectively.

6 Conclusion and future work

In this paper, we propose the ChNWI approach to identify Chinese new words of three schemas. We first summarize three schemas based on eight surfaces, they are single-character schema (covers NW11, NW111 and NW1111), affix schema (spans NW21, NW31, NW12 and NW13) and NW22 schema. Next, we represent that four linguistics features, i.e., *word frequency*, *co-occurrence probability*, *mutual information* and *adjacent category*, play same impacts on the three schemas, while *independent word probability* is important to single-character schema, *head-character word probability* and *tail-character word probability* are key facts to prefix schema and suffix schema respectively. Our experimental results on two corpora show that, new words are categorized into three schemas and employing their unique features not only improve the accuracy score but also improve the recall rate of identification.

We also test the ChNWI approach on the domain-related (Mobile Communication) corpus with 80 thousand sentences. All these sentences are collected from Baidubaike and Wikipedia. With the development of new business in the Mobile Communication domain, there are a considerable amount of new words which are not registered in the dictionary of a segmenter. We test our approach in identifying new words of the three schemas contained in the domain-related corpus. The ChNWI approach gets three accuracy rates, 80%, 68% and 71%, for single-character schema, affix schema and NW22 schema respectively.

In future, we further improve the ChNWI approach from the following three aspects: (1) apply automatic feature selection and check the performance; (2) consider the combination of different schemas for other surfaces (i.e., NW211 and NW112). (3) study additional schemas rather than the three suggested schemas.

Acknowledgement

This work is sponsored by the grant from the Shanghai Science and Technology Foundation (No. 11511504002).

References

Chih-Chung Chang, and Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*.

Guodong Zhou. 2005. A chunking strategy towards unknown word detection in Chinese word segmentation. *Lecture Notes in Computer Science*, 3651:530–541.

Haijun Zhang, Shumin Shi, Chaoyong Zhu, and Heyan Huang. 2010. Survey of Chinese New Words Identification. *Computer Science*, 37(3):6–11.

Hongqiao Li, Chang-Ning Huang, Jianfeng Gao, and Xiaozhou Fan. 2005. The use of SVM for Chinese new word identification. *First international joint conference on Natural Language Processing*, 723–732.

Huiming Duan, Zhifang Sui, Ye Tian, and Wenjie Li. 2012. The CIPS-SIGHAN CLP 2012 ChineseWord Segmentation onMicroBlog Corpora Bakeoff. *Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 35–40.

Jiahen Zheng, and Wenhua Li. 2002. A Study on Automatic Identification for Internet New Words According to word-building Rule. *Journal of Shanxi University (Natural Science Edition)*, 25(2):115–119.

Kaixu Zhang, Maosong Sun, and Changle Zhou. 2012. Word Segmentation on Chinese Mirco-Blog Data with a Linear-Time Incremental Model. *Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 41–46.

Longye Wang, Derek F. Wong, Lidia S. Chao, and Junwen Xing. 2012. CRFs-Based Chinese Word Segmentation for Micro-Blog with Small-Scale Data. *Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 51–57.

Ning Xi, Bin Li, Guangchao Tang, Shujian Huang, Yinggong Zhao, Hao Zhou, Xinyu Dai, and Jiajun

Chen. 2012. Adapting Conventional ChineseWord Segmenter for Segmenting Micro-blog Text. *Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 63–68.

Shengfen Luo, and Maosong Sun. 2003. Two character Chinese word extraction based on hybrid of internal and contextual measure. *Second SIGHAN Workshop on Chinese Language Processing*, 24–30.

Tao-Hsing Chang, and Chia-Hoang Lee. 2003. Automatic Chinese unknown word extraction using small-corpus-based methodmeasure. *1st IEEE International Conference on Natural Language Processing and Knowledge Engineering*, 459-464.

Wenbo Pang, Xiaozhong Fan, Yijun Gu, and Jiangde Yu. 2009. Chinese Unknown Words Extraction Based on Word Level Characteristics. *9th International Conference on Hybrid Intelligent System*, 361–366.

Xiao Sun, Degen Huang, Haiyu Song, and Fuji Ren. 2011. Chinese new word identification: a latent discriminative model with global features. *Journal of Computer Science and Technology*, 26(1):14–24.

Xin Jiang, Yanjiao Cao, and Zhao Lu. 2011. Automatic Recognition of Chinese Unknown Word for Single-Character and Affix Models. *Sixth International Conference on Intelligent Systems and Knowledge Engineering*, 435–444.

Yisu Xu, Xuan Wang, Buzhou Tang, and Xiaolong Wang. 2008. Chinese Unknown Word Recognition using improved Conditional Random Fields. *8th International Conference on Intelligent Systems Design and Applications*, 363–367.

Ziru Zhang, Qiangjun Wang, and Xuedong Tian. 2006. Chinese New Words Extraction Based on Machine Learning Approach. *2006 International Conference on Machine Learning and Cybernetics*, 3380–3384.